DOCUMENT RESUME

ED 341 729 TM 017 924

AUTHOR Naizer, Gilbert
TITLE Basic Concepts in Generalizability Theory: A More Powerful Approach to Evaluating Reliability.
PUB DATE Jan 92
NOTE 19p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Houston, TX, January-February 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Error of Measurement; *Estimation (Mathematics); *Generalizability Theory; Higher Education; Interrater Reliability; Measurement Techniques; *Research Design; Research Methcdology; Test Interpretation; *Test Reliability; Test Theory

ABSTRACT

A measurement approach called generalizability theory (G-theory) is an important alternative to the more familiar classical measurement theory that yields less useful coefficients such as alpha or the KR-20 coefficient. G-theory is a theory about the dependability of behavioral measurements that allows the simultaneous estimation of multiple sources of error variance. If error influences interact, as they often will, the G-theory estimates may be markedly different from classical theory estimates. G-theory also distinguishes between relative and absolute decisions. Finally G-theory provides a mechanism for using estimated error variances for alternative designs (D-studies) to help researchers develop a measurement that minimizes error for a future study, but that is also efficient. Some of the major advantages of G-theory are explained and illustrated with a hypothetical study of 20 individuals given a performance task on 3 occasions and assessed by 2 raters. Three tables present data from the example. A five-item list of references is included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

g-theory

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

GILBERT NAIZER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED341729

Basic Concepts in Generalizability Theory: A More Powerful Approach to Evaluating Reliability

Gilbert Naizer
Texas A&M University 77843-4232

Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX, January 31, 1992.

TM017924

BEST COPY AVAILABLE

ABSTRACT

An important measurement approach called generalizability theory is an important alternative to the more familiar classical measurement theory that yields less useful coefficients such as alpha, or the KR-20 coefficient. G-theory allows the simultaneous estimation of multiple sources of error variance. If error influences interact, as they often will, then G-theory estimates may be markedly different from classical theory estimates. G-theory also distinguishes between relative and absolute decisions. Finally, G-theory provides a mechanism for using estimated error variances for alternative designs (D-studies) to help the researcher develop a measurement that minimizes error for a future study but that is also efficient. The present paper explains generalizability theory and some of its several major advantages.

Generalizability theory (G-theory) is a theory about the dependability of behavioral measurements. Dependability in this sense is how accurately we can generalize from a person's observed score on a test or other measure to the average score that person would have received under all possible conditions that the test user would be willing to accept. When the focus of measurement is on presumed true differences in individuals, we assume that the differences among an individual's scores under different conditions are due to one or more sources of measurement error, and not to changes in the individual. But a single person's score on one occasion is not fully dependable. The score would usually be different on other occasions, on other test forms, or with different test administrators. G-theory provides a means for determining and making decisions dealing with the dependability of a measurement.

Although G-theory is a very powerful approach to assessing reliability, most researchers are more familiar with the more traditional theory, called classical theory, which yields coefficients such as KR-20 or test-retest reliability. The present paper explains the much more useful G-theory and its potential benefits.

Advantages of G-Theory over Classical Theory

Shavelson, Webb, and Rowley (1989) use the analogy of comparing simple ANOVA to factorial ANOVA to indicate some of the differences between classical theory and G-theory. Some of these fundamental differences between classical theory and G-theory have

been noted by Shavelson et al.:

> The concept of reliability, so fundamental to classical theory, is replaced by the broader and more flexible notion of generalizability. Instead of asking how accurately observed scores reflect their corresponding true scores, generalizability theory asks how accurately observed scores permit us to generalize about a persons' behavior in a defined universe of situations. (p. 922)

Classical theory considers only one source of variation *at a time*. Test-retest reliability considers the occasion as the source of error, internal consistency reliability considers the item as the source of error, and parallel forms reliability considers the form of the test as the source of error (Webb, Rowley & Shavelson, 1988). As Thompson (1991) emphasizes, most measurement classicists unconsciously presume that these error sources overlap and also do not interact to create additional new error variance.

A researcher may compute all three of the previously mentioned classical theory reliabilities and find that in all three measurements error comprises 10% of the score variance. Many researchers would assume that the ten percents are the same and also not realize that an interaction between two or more of these could readily create completely different additional sources of measurement error.

In contrast, G-theory allows the estimation of the magnitude of multiple sources of error *simultaneously*. G-theory also

estimates interactive sources of measurement error not considered in classical theory. These estimations allow the researcher to determine to what extent the results of a measurement are generalizable to a population, occasion, or other administrator, etc., using what are called generalizability studies (G-studies).

In addition, using decision or D-studies, G-theory enables the decision maker to use the G-study results to determine how many occasions, administrators, test forms, etc. are required to obtain dependable scores in the most efficient manner. For example, D-studies may help a researcher decide "What measurement protocols can I use to get a generalizability coefficient of at least .85?" and, given this answer to this question, the researcher can also decide which of the competing acceptable protocols is cheapest, least intrusive, and so forth.

Although G-theory provides a generalizability coefficient, the theory focuses on the "variance components" that index the magnitude of each source of error (Shavelson, Webb, and Rowley, 1989), and that are the actual basis for the calculation of generalizability coefficients. These variance components allow the researcher to determine the major sources of error variance when making decisions or generalizations.

Another strength of G-theory is the differentiation of "relative" and "absolute" decisions. Relative decisions concern the rank ordering of individuals, such as percentile ranking on achievement tests, and are not concerned with the actual score a person receives. Absolute decisions are based on the absolute

level of performance (actual score) of an individual. For example, a certification test with a minimum passing score depends on the number of items answered correctly, and not solely on how the individual performed in relation to others taking the test. The dependability of the scores for relative and absolute decisions may not be the same, as will be explained shortly.

Multiple Sources of Error Variance

The foundation of G-theory is the definition of a universe of admissible observations. This universe consists of all the observations that the researcher is willing to treat as interchangeable (e.g., a score on a different occasion) for the purposes of making a decision. Within this universe are aspects of measurement called facets. A facet is a single source of measurement, with the levels of the facet (conditions) usually being assumed to be infinitely large. For example, if the researcher wishes to generalize from one test form to a larger set of test forms, FORMS is a facet and the levels of the facet are all "admissible" (i.e., acceptable) test forms. Measurement error is present whenever a generalization from a particular measurement to behavior within the larger universe is made.

The universe can be single-faceted, with the researcher intending to generalize within only one source of error variation (e.g., forms), or multi-faceted where the researcher intends to generalize across several sources of error variation (e.g., forms, occasions, and administrators). As an example of a single-facet universe, consider an achievement test in which the students'

achievement scores are based on the sample of items which is on the test. The items universe consists of all possible achievement test items of which those on the test are only a sample. Ideally, the researcher wants to know each person's universe score; however, we must use the test score to generalize from this particular set of items to the student's "universe" of achievement scores.

This one-facet design has four sources of variability. These variability sources are due to (a) systematic differences among students (called the object of measurement), (b) differences in item difficulty, (c) person x item interaction (some items are easier for some students), and (d) random or unknown events. The 3rd and 4th variability sources cannot be separated and are lumped together in a residual variation. The object of measurement is the object about which the researcher wishes to generalize (usually persons) and therefore by definition creates what is considered systematic variance. The facets, in this case items, contain error variance. The systematic variance is due to differences in the object of measurement presumed to be real, and is therefore desirable, while the variances from the facets and their interactions are presumed to not be real and to therefore be measurement error.

A performance assessment task on which the students are rated by multiple raters on multiple occasions represents a two-facet design, with items and occasions as the facets. This design would have seven sources of variability--one systematic source for persons and six error sources--as reflected in Table 1. The six

error variability sources consist of: (a) differences in raters, (b) differences in occasion, (c) person x rater interaction (raters may rate some students harder than others), (d) person x occasion interaction (some students may do better on one occasion), (e) rater x occasion interaction (raters may grade easier on one occasion), and (f) residual--person x rater x occasion interaction combined together with unmeasured or random events.

INSERT TABLE 1 ABOUT HERE.

Designs can obviously contain more than two facets with many sources of variability. In any G-study, it is important that the facets the researcher wishes to generalize over be included in the study (Webb, Rowley, & Shavelson, 1988) and it should be noted that the broader the universe of admissible observations, the greater is the possibility of making an error in generalizing from the sample to the universe (Shavelson & Webb, 1991, p. 10).

Designs can also be "nested", as when different test forms consist of different sets of items (Shavelson & Webb, 1991). Nested designs and fixed facets will not be covered in this paper, but for a good discussion of these topics one can consult Shavelson and Webb (1991).

Variance Components

Each source of variability from a study has an associated variance called the "variance component". These variance components are the focus of G-theory and can be estimated using the Expected Mean Square (EMS) equations of the ANOVA procedure. The statistical model and mathematical treatment for variance components in G-theory can be found in Shavelson and Webb (1991). Any analyses of variance computer program (SAS, SPSS, BMDP8V) can be used to obtain the estimated variance components. In addition, Crick and Brennan's GENOVA program has been developed specifically for generalizability theory, and finds EMS equations, estimated variance components, and generalizability coefficients.

Consider the previously mentioned two facet design. This hypothetical example will be used to present a numerical example in which 20 individuals were administered a performance task on three occasions. Two raters assessed the students performance on all occasions. This situation represents a fully crossed design with persons crossed with raters and occasions and is denoted as p x r x o.

Table 2 gives the estimated variance components and the percentage of the total variance for each. Negative estimated variance components can occur, although negative variance is conceptually impossible, since scores can never be less "spread out" than not spread out at all. Negative estimates can arise because of misspecification of the model or because of sampling error (Shavelson & Webb, 1991). Several methods of dealing with negative variance components have been developed (Shavelson, Webb & Rowley, 1989). Our example is simplified because it does not contain negative estimates.

INSERT TABLE 2 ABOUT HERE.

The variance component for persons (.3645 from Table 2) accounts for 39% of the total variance. This indicates that persons systematically differed somewhat in their performance; this is desirable variability since persons are the object of measurement, i.e., persons are presumed to have legitimately different scores. The next largest component, the residual (24%) indicates that a large portion of the variance is due to *either* the three-way interaction of the facets or variation sources that were not measured in the study. The large component due to raters (11%) is disturbing, indicating substantial disagreement across raters as regards the performance ratings of the students. The high rater x occasion component indicates an inconsistency in raters' ratings on different occasions. The variance component for persons x raters shows that raters disagreed somewhat on the relative performance of the students. These high variances due to raters and rater interactions (29% of the total without including the residual) may indicate that additional training for raters or a improved rating system is needed.

The relatively low person x occasion variance component (5%) indicates that the relative performance of students did not vary greatly from occasion to occasion. The smallest component, occasion (3%), indicates that student performance did not differ much by occasion. This result suggests that measuring performance repeatedly at different times would not yield much improvement in

measurement integrity.

The results from our example indicate that a larger portion of the variability can be attributed to raters than to occasion. The following section follows this example and describes how to use the variance components to determine the dependability of scores.

Generalizability Coefficients

The variance components from the example can be used to calculate classical reliability coefficients, but in G-theory, two types of coefficients can be calculated and used in D-studies. These coefficients, one for relative decisions (i.e., decisions based on stability of ranks ignoring other consideration) and one for absolute decisions (i.e., decisions against an absolute standard such as a number-of-right-answers fixed criterion), are calculated differently. For relative decisions, the error variance is the sum of only the subset of components that affect the relative standing of individuals. For a crossed design with, the components that affect relative standing are those interactions containing the object of measurement. In our example, p x r, p x o, and p x r x o,e are included. The main effects for raters and occasions and the interaction between these two are ignored when evaluating a relative decision since they do not contribute to the relative standing of persons. For absolute decisions, on the other hand, the error variance is the sum of all variance components except that for the object of measurement itself.

For a given study, the researcher may be interested in either a relative or absolute decision, or both. For purposes of

demonstration, both will be discussed in this example. The generalizability coefficient (ρ^2_{rel}) for relative studies (analogous to the reliability coefficient in classical theory) is systematic score variance divided by expected score variance. This is the variance for the object of measurement divided by the object of measurement variance plus the relative error variance ($\sigma^2_p/(\sigma^2_p+\sigma^2_{rel})$). The relative decision error variance (σ^2_{rel}) is a summation of all interactions involving persons with each divided by the number of conditions in each facet (.0467/3 + .0748/2 + .2243/2x3), i.e., .0904. Using this value, the generalizability coefficient (ρ^2_{rel}) then is .3645/(.3645 + .0904) = .801.

For an absolute decision, all variance components except the universe score variance are included in the summation. The absolute error variance (σ^2_{abs}) for this example (.0128/2 + .0280/3 + .0748/2 + .0467/3 + .0935/2x3 + .2243/2x3) is .1667. A reliability-like coefficient for absolute decisions (ϕ, phi) can be calculated similar to the generalizability coefficient, using the formula $\phi = (\sigma^2_p/(\sigma^2_p+\sigma^2_{abs}))$. This coefficient is not actually a generalizability coefficient since the denominator is not the observed-score variance and the coefficient does not approximate the expected value of the squared correlation between observed and universe score scores (Shavelson & Webb, 1991). The phi coefficient for our example is .3645/(.3645 + .1667) = .686.

D-Studies

Decision studies allow the researcher to use the results from the G-study to determine the number of occasions, raters, tests,

etc. required for dependable scores. D-studies allow the projection of alternative measurement designs by varying the number of conditions of each facet and calculating variance components and generalizability coefficients for each of the alternative designs. The information from these alternative designs assists the decision maker in designing future measurements. The alternative designs and other information (e.g., relative costs of adding a rater or occasion) are used concurrently in decisions about future studies. Eason (1991) points out that a D-study cannot include facets that were not included in the G-study, although a facet may be eliminated.

D-study information from six alternative designs of our example are presented in Table 3. Since raters create more error variance than occasions, we expect that changing the number of raters will have a greater effect on the generalizability and phi coefficients than changing the number of occasions. From Table 3 we can see that adding 1 rater increases the generalizability coefficient from .801 to .882 while adding 1 occasion causes a change from .801 to .844 (phi coefficients behave similarly and for simplicity will not be discussed here). Furthermore, doubling the number of occasions to six does not produce as large an increase (.801 to .878) as adding 1 rater. Perhaps an increase of one in both occasion and rater produces the desired generalizability coefficient level (.913).

The researcher determines the desired level for the generalizability and phi coefficients. Then the researcher asks a

series of "what if" questions exploring the effects of adding or subtracting levels from the facets, until acceptable coefficients are realized from one or more protocol. It should be kept in mind that the improvement from adding additional levels diminish with each addition. For example, adding the fourth rater produces only 40% of the effect of adding the third rater. Once several protocols that yield acceptable coefficients are isolated, the researcher selects the protocol that is most efficient or least cumbersome.

Summary

G-theory extends classical theory in several ways. The theory allows the simultaneous estimation of multiple sources of error variance. If error influences interact, as they often will, then G-theory estimates may be markedly different from classical theory estimates (Eason, 1991). G-theory distinguishes between relative and absolute decisions and provides a generalizability estimate for each type of decision. G-theory provides a mechanism for using estimated error variances for alternative designs (D-studies) to help the researcher develop a measurement that minimizes error for a future study but that is also efficient. G-theory can be implemented using several computer software packages (SAS, SPSS, BMDP8V, GENOVA) and applied to a wide variety of designs.

According to Shavelson, Webb and Rowley (1989), G-theory provides perhaps the most flexible measurement theory available to psychologists. Thompson (1991, p. 1072) suggests that:

> [T]oo few researchers recognize that in all analyses we inherently invoke both a presumptive model of reality and an analytic model. When the two don't

> match, the analysis doesn't help us understand the reality we believe exists. If we virtually always want to generalize over time and over items or tests, then a classical theory approach that never simultaneously considers these two time and item sampling influences, and completely ignores the interactions of these influences, will be quite simply unworkable!

The use of G-theory will increase if researchers become aware of the important notion emphasized by Eason (1991): "only G-theory honors a complex reality in which measurement error sources may interact to compound each other!"

References

Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples. In B. Thompson (Ed.), Advances in Educational Research (Vol. 1, pp. 83-98). Greenwich, CT: JAI Press.

Shavelson, R.J., & Webb, N.M. (1991). Generalizability theory: A primer. Newbury Park, CA: SAGE Publications.

Shavelson, R.J., Webb, N.M., & Rowley, G.L. (1989). Generalizability theory. American Psychologist, 44, 922-932.

Thompson, B. (1991). Review of Generalizability theory: A primer by R.J. Shavelson & N.W. Webb. Educational and Psychological Measurement, 51, 1069-1075.

Webb, N.M., Rowley, G.L., & Shavelson, R.J. (1988). Using generalizability theory in counseling and development. Measurement and Evaluation in Counseling and Development, 21, 81-90.

Table 1
Sources of Variability in a Two-facet Design

Source of Variability	Type of Variation	Notation
persons (p)	universe-score	σ^2_p
raters (r)	rater difficulty	σ^2_r
occasion (o)	occasion difference	σ^2_o
p x r	interaction	σ^2_{pr}
p x o	interaction	σ^2_{po}
r x o	interaction	σ^2_{ro}
p x r x o, e	residual	$\sigma^2_{pro,e}$

Table 2
Estimated Variance Components for P X R X O Design

Source	Estimated Variance Component	Percent
persons	.3645	39
raters	.1028	11
occasion	.0280	3
pr	.0748	8
po	.0467	5
ro	.0935	10
pro,e	.2243	24

Table 3
Alternative Design Variance Components and
Generalizability Coefficients

	G-study	D-study					
n_r	2	2	2	3	3	4	4
n_o	3	4	6	3	4	3	5
p	.3645	.3645	.3645	.3645	.3645	.3645	.3645
r	.1028	.1028	.1028	.0685	.0685	.0514	.0514
o	.0280	.0210	.0140	.0280	.0210	.0280	.0168
pr	.0748	.0748	.0748	.0499	.0499	.0374	.0374
po	.0467	.0350	.0234	.0467	.0350	.0467	.0280
ro	.0935	.0701	.0468	.0623	.0468	.0468	.0281
pro,e	.2243	.1682	.1122	.1495	.1122	.1122	.0337
σ^2_{rel}	.0857	.0672	.0506	.0488	.0347	.0343	.0166
σ^2_{abs}	.1667	.1316	.1083	.0900	.0667	.0604	.0469
ρ^2	.801	.844	.878	.882	.913	.914	.956
ϕ	.771	.735	.771	.802	.839	.858	.886